Using Linear Visualization to Explore Large Graphs

William J.R. Longabaugh

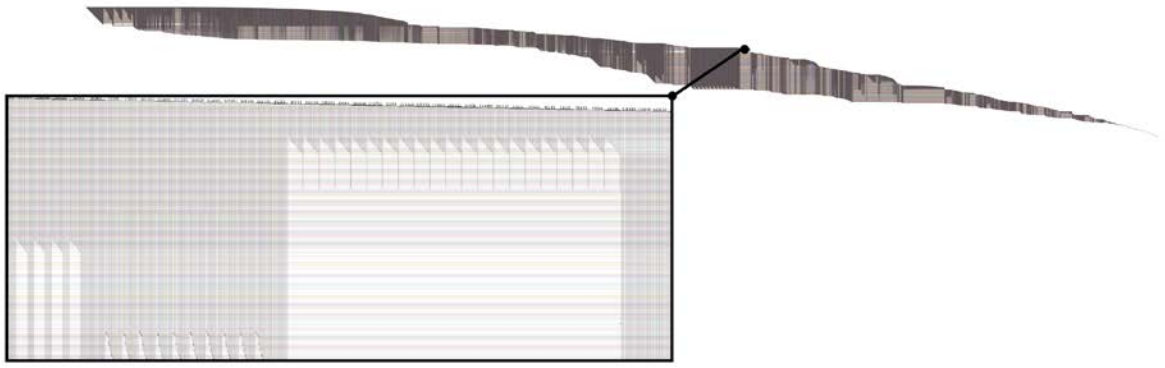

Fig. 1: The Stanford Web network (281,903 nodes, 2,312,497 edges) [1] shown using BioFabric. Inset shows detail from top edge.

Abstract—Traditional node-link diagrams simply do not scale, primarily due to the rarely-questioned convention that nodes must be represented as fundamentally zero-dimensional points in 2D or 3D space. The software tool BioFabric [2] abandons this nodes-as-points convention, and instead depicts nodes as *horizontal lines* and edges as *vertical lines*. This completely decouples node and edge placement, thus leading to a scalable visualization technique. Significantly, this approach creates a node-link diagram where the visual and logical representation of the network is transformed into a fundamentally linear, sequential construct. Such a model is an natural choice for visualizing graphs at scale.

Index Terms—Graph visualization, node-link diagrams, scalability issues

1 Introduction

Despite the fact that graph sizes have grown incredibly in recent years, the techniques used visualize them have remained basically static; we still rely on traditional node-link diagrams and adjacency matrices, regardless of how big the graph is or how useful the resulting visualization is for these huge networks.

Consider the use of dimensions in these traditional methods. Typical node-link diagrams represent nodes as zero-dimensional points and links as one-dimensional lines, both depicted in a shared (overloaded and crowded) two- (or three-) dimensional space; assuming straight lines are used for the edges, node placement *fully* constrains edge placement. Alternatively, adjacency matrices for a graph with n nodes and e edges represent edges as zero-dimensional points placed on an n x n grid arranged on the two-dimensional plane, where typically $n \times n >> e$; thus most of the plane is wasted for sparse graphs.

To attack the problem of visualizing graphs at scale, it is worthwhile to think about how to maximally leverage the limited number of dimensions we have at our disposal. To drive this discussion, I will describe BioFabric [2, and also the topic of a poster here at IEEE Vis 2015], which depicts nodes not as zero-dimensional points, but as one-dimensional *horizontal lines*; edges are *vertical lines* (see Figure 1). In addition to removing all ambiguity in the visual presentation of the network, an important outcome of this representation is that the mental model of a huge network becomes a linear sequence, which is a natural way for people to organize complex information.

- *William J.R. Longabaugh is with the Institute for Systems Biology. E-mail: wlongabaugh@systemsbiology.org.*

(Manuscript and copyright information block).

2 Previous Work

There have been many recent efforts to incrementally improve node-link diagrams, notably hive plots [3], edge bundling [4][5], motif simplification [6], power graphs [7], and others. But it seems unlikely that these approaches can be leveraged to handle truly huge graphs. To tackle that problem, a fundamental change in representation seems to be a better approach.

2.1 "Nodes as Lines"

Lines and linear shapes have previously been used to depict nodes in node-link diagrams. Using finite-length rectangular representations for nodes has been the domain of visibility representations (VRs) for planar graphs [8, 9] for many years, though it is significant that VRs have been *specifically* focused on handling the case where edges are constrained to *not* intersect any of these horizontal node regions. Blakley [10] went a step further and allowed node-link intersections by describing an algorithm for nonplanar graphs which uses horizontal "bricks" for nodes and depicts edges "tunnelling under" the bricks using a special symbol.

McAllister used the nodes-as-lines technique in a figure [11] to illustrate an algorithm for the linear arrangement problem; this was a natural representation of the domain that allows the edges to be clearly shown despite the one-dimensional nature of the problem.

A currently common nodes-as-lines usage is in Unified Modeling Language (UML) sequence diagrams, where objects have an associated vertical lifeline [12]. Notably in that context, node lines are being used to represent objects through the passage of time; people seem to find this specific use of nodes as lines as fundamentally intuitive.

Finally, it is notable that a node-link diagram with a row-and-column organization starts to blur into the domain of matrix visualizations. For example, in an incidence matrix, each node is a

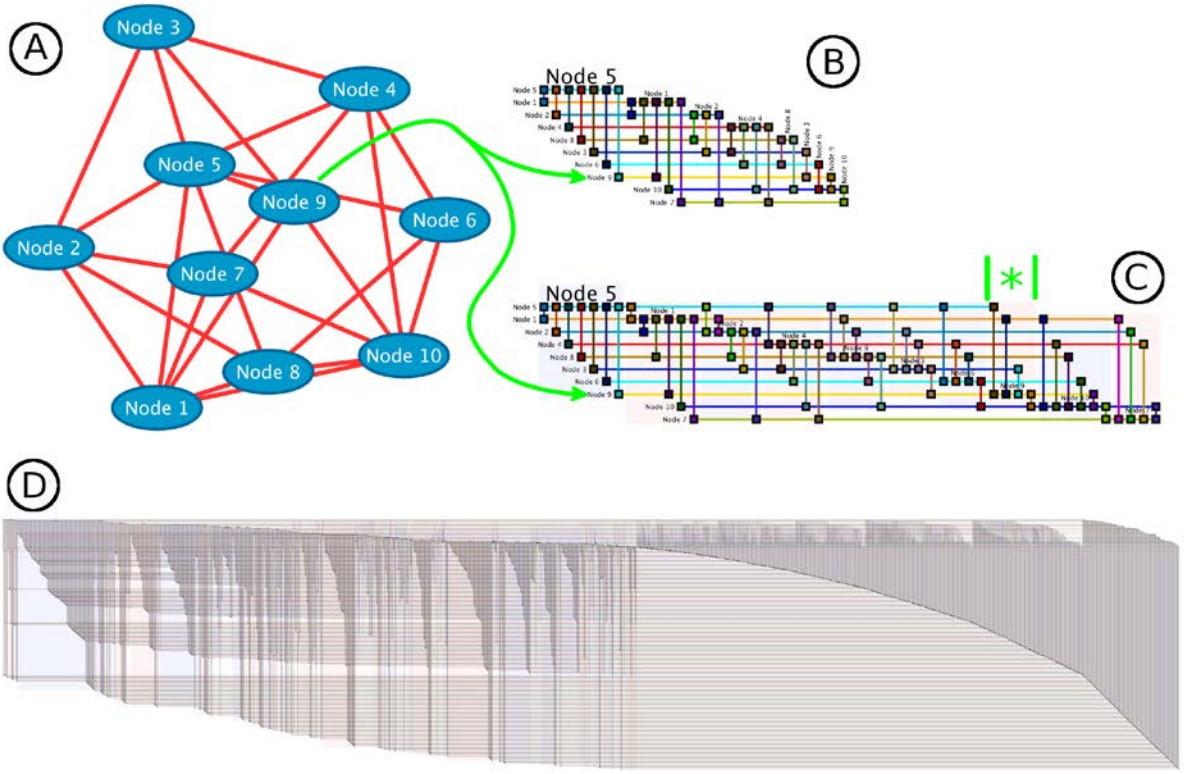

Fig. 2: A conventional node-link diagram **(A)** is shown in **(B)** using BioFabric, and shown in **(C)** using *shadow links*. The green asterisk in **(C)** highlights how all edges incident on *Node 9* can be seen in one place; compare this to **(B)**. **(D)** shows how shadow links can be used to visualize target structure in the E Coli gene regulatory network from [14]. That network has 1,793 nodes and 4,268 edges; shadow links double the final edge count.

row and each edge is column; the ones in a column represent the endpoints of that edge. BioFabric can be thought of as a node-link diagram re-interpretation of this overall planar organization. In light of this correspondence, BioFabric shares a structural and visual similarity to the matrix-based approach Genaquilts [13]. Genaquilts uses matrix rows to represent nodes (people), and columns to represent parent-child relationships; i.e. the columns can be viewed as representing hyperedges, and thus Genaquilts is depicting an incidence matrix of a hypergraph.

3 BioFabric

BioFabric grew out of the idea that drawing nodes as parallel lines, instead of as points, could provide distinct advantages for representing a node-link diagram; compare Figures 2A and 2B. With nodes arranged as parallel horizontal lines, one per row, edges are then presented as vertical lines, drawn in front of the nodes, each arranged in a unique column. The two endpoints of each edge are marked with a small square glyph, and the edges are drawn darker than the nodes.

Of course, this means that a BioFabric network by its very nature is full of line intersections; even a small network display will contain easily over a million node/edge line intersections (though there are exactly zero edge intersections). The trick is that only intersections with the glyph symbol have any meaning, and all the rest can be ignored. Also the uniformly darker links drawn in front of the uniformly lighter nodes makes the links stand out as well, reducing the visual impact of the intersections. In effect, intersections are so common, regular (i.e. strictly orthogonal), unremarkable, and unambiguous that they are easily ignored, as only glyph-marked intersections are of interest.

Finally, colors are used in a fixed, repeating cycle to help the user maintain context while scanning for a long distance along a node or edge line; compare this with trying to trace a single black line among a vast expanse of other black line. While this means that color is not available to represent edge attributes, I argue that graph readability must take first priority.

One important feature to note is that the standard layout for edges creates contiguous, ordered, parallel edge sets for each node, thereby creating a distinct, characteristic, and visually prominent *edge wedge*. The shape of the wedge visually conveys how that node is connected within the network. Compare this to an adjacency matrix, where that information is instead collapsed into a single column of points. Significantly, these edge wedges can be directly compared between multiple nodes to see how similar or different their node neighborhoods are. As an example, see the prominent wedges on the left half of Figure 2D.

3.1 Shadow Links

By default, BioFabric displays only one copy of each link in the network, directly corresponding to the standard node-link representation. But when *shadow links* are enabled, *two* copies of each link are drawn; compare Figures 2B and 2C. At the expense of doubling the width of the diagram, this simple duplication step makes it possible to show, in a single compact region on every node line, *all* the incident links for that node. This is particularly useful for large networks, e.g. Figure 2D, and is also invaluable for allowing direct visual comparison of the edge wedges associated with two or more nodes or link groups, described next.

3.2 Link Groups and Network Comparisons

Another valuable feature is *link groups*. The default edge layout algorithm creates a single edge wedge for each node, but link groups allow the user to easily reorganize this single wedge into multiple adjacent wedges, based upon some attribute. Thus, for example, the edges in a multigraph can be easily separated into distinct components by assigning the different edge types to different link groups, creating different wedges.

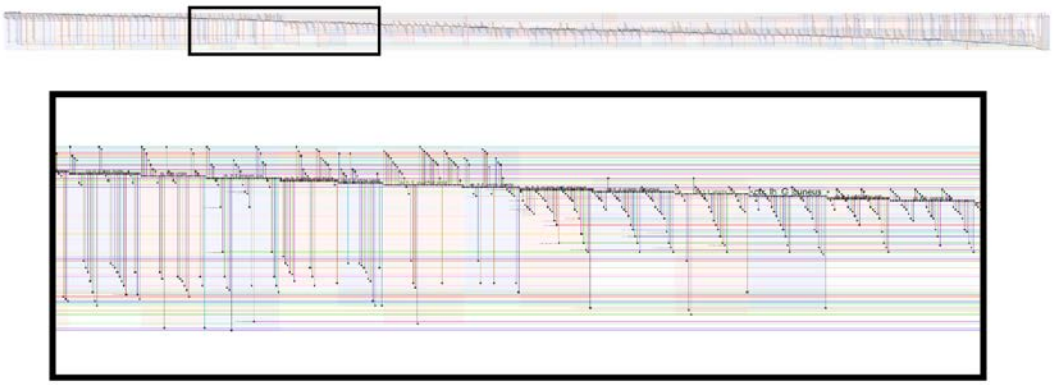

Fig. 3: A shadow link and link-group based network comparison of three resting-state fMRI networks; data from [15][16]. The top image is the full network view, while the highlighted detail shown below illustrates how each node (one for every shaded pink or blue zone) has three distinct edge wedges, allowing the user to quickly pan across the network and assess the differences between the networks.

Finally, two or more large networks can be visually compared in a systematic fashion by using link groups to separate the edges belonging to each network into distinct edge wedges. Figure 3 shows such a comparison between three separate networks

3.3 Is BioFabric Intuitive?

A frequent criticism of nodes-as-lines is that it is "not intuitive", and given the way people have been trained to think that nodes are represented by small point-like geometric shapes, this is perhaps unsurprising. Anecdotally, some people catch on to the notion almost immediately, while others need to be walked through the logic and take some time to understand it. In all cases, it takes experience to be able to usefully interpret a BioFabric presentation and use its unique features to glean interesting insights about the network from it.

But I will argue that this sometimes-steep learning curve is a worthwhile investment of effort for users who are professionally involved in large network analysis and interpretation. The need to be trained to interpret these diagrams should not cause it to be ignored in favour of existing highly "intuitive" but ineffective techniques. The best way to approach training to interpret the diagrams is an interesting open question; perhaps the insight that people seem to easily understand nodes as lines when they are used to show passage of time (as with UML sequence diagrams) provides a basis for introducing the concept.

4 New Challenges: Going for Scale

The released Version 1.0 of BioFabric has served as a proof of concept to demonstrate the feasibility and advantages of nodes as lines. Although the current implementation has some features to accommodate large graphs, such as a cached image tile system for rendering the global small-scale network views quickly, much work remains to be done. For example, the graph shown in Figure 1, with 2.3 million edges, is already demonstrating very marginal performance on the existing platform. However, there does not appear to be any inherent limitations for scaling the basic nodes-as-lines abstraction to truly large networks. In this section, I will discuss some of these issues.

4.1 The Network as a Linear Entity

As mentioned above, an important outcome of this representation is that the user's mental model of a huge network becomes a *simple linear sequence*, which is a great advantage when dealing with extremely large data sets. Small subsets of the network that maintain this overall linear mental model can be created simply by cutting them out and pasting them back together using the same global order. Also, while the current implementation primarily deals with horizontal navigation simply by using the scrollbar, a hierarchical system of bookmarking and indexing could be used to quickly access short pieces of the full network "strip" for very large models. For extremely long and thin networks (i.e. those with high average node degree), presenting the network in the manner of a musical score (horizontally stacked segments of fixed width) might be a fruitful approach to investigate.

4.2 The Vertical Dimension

Perhaps more challenging is visualizing the vertical dimension of an extremely large network. Consider the situation in Figure 4, which shows the transitive reduction, with 62,257 edges and 27,383 nodes, of an academic citation network (data from [17]). The top panel shows the full network (with shadow links, so the network width is doubled), while the frame on the lower left shows the zoomed-in view of the diagonal located at the green crosshair. Unsurprisingly, a lot of structure can be gleaned just by looking near the diagonal, since many links go to nearby neighbors. But many links disappear out of the frame as well, and if the network contains a billion nodes, this near-the-diagonal view would be of limited use.

But the lower right panel of Figure 4 shows one approach, which is to only show the actual node neighbors in the view. All the other nodes are not drawn, and these unused rows are compressed out of the display, creating the small but complete view shown in the figure that faithfully maintains the global ordering of the nodes. Note that if node rows have been organized at a high level into contiguous annotated groups, visual remnants of these group annotations could also be retained in this compressed view, even if none of their nodes were present. This high-level information would provide additional context for how the compressed view relates back to the model as a whole.

With this eliminate-and-compress approach, if the user is trying to view a set of x nodes along the diagonal, and the average node degree in the network is y, then the number of displayed node rows will typically be between y and xy, depending on the similarity of the node neighborhoods among the x nodes being explored. If this number is still too large to accommodate on the display, it could be augmented by the option of introducing a vertical 1-hyperbolic projection technique, e.g. [18].

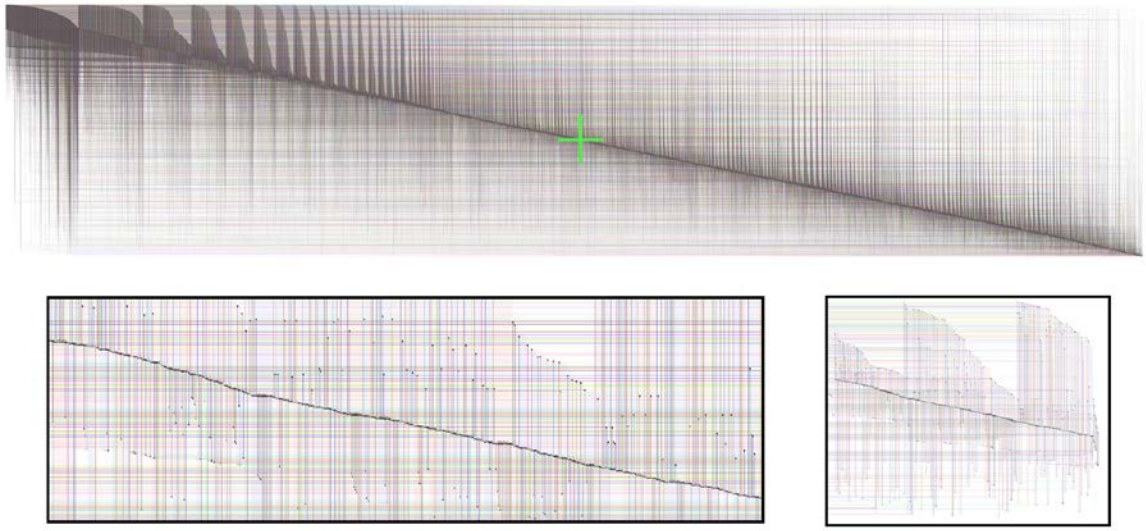

Fig. 4: Transitive reduction of the citation network of arXiv high energy theory papers; data from [17]. This network has 62,257 edges (double that for shadows) and 27,383 nodes. Upper panel is the full network. Lower left show the view close to the diagonal; not unexpectedly, links disappear off the screen above and below. Lower right suggests an alternative view; nodes that do not have incident edges are removed and the view is vertically compressed. This approach could be combined with a vertical 1-Hyperbolic projection [18] for very large networks.

4.3 The Global View and Zooming

Even with very large networks, the global view can provide useful information. For example, the aspect ratio of the *n* x *e* global view instantly gives the user a visual cue of the average node degree in the network. Furthermore, the overview provides a context in which to begin drilling down into a network, and providing a global sense of place. For example, the prominent contiguous groupings of similar nodes in Figure 1 provide a global pattern that is somewhat reminiscent of the navigational framework provided by chromosome banding.

Interactive zooming from the global level down to the smallest detail is currently possible in BioFabric using its image tile cache. This feature invites the user to browse at will through the network to look for interesting patterns, but the usefulness or practicality of this approach for networks with millions of nodes is an open question. It seems likely that at certain levels of detail, some large-scale structures would emerge for a given layout that would be useful to pre-compute and represent at least approximately in a visualization, though most of the network landscape between the two extremes of scale is probably unremarkable and of limited value.

5 Conclusion

To create useful visualizations of huge networks, it is necessary to completely re-evaluate both the fundamental allocation of the limited set of available dimensions, as well as the dimensionality of the entities used to represent nodes and edges.

Acknowledgments

BioFabric was developed with assistance from NCI award U24CA143835. The author is currently supported by NICHD grant R01HD073113. Data for figure 3 were provided [in part] by the Human Connectome Project, WU-Minn Consortium; see [19].

References

[1] J. Leskovec, "Stanford web graph", *SNAP: Network datasets*, http://snap.stanford.edu/data/web-Stanford.html, 2002.

[2] W.J.R. Longabaugh. Combing the hairball with BioFabric: a new approach for visualization of large networks. BMC Bioinformatics, 13:275, 2012.

[3] M. Krzywinski, et al. Hive Plots - Rational Approach to Visualizing Networks. Brief Bioinform 13(5): 627-644, 2012.

[4] D. Holten. Hierarchical edge bundles: visualization of adjacency relations in hierarchical data. IEEE Trans Vis Comput Graph, 12(5): 741-8, 2006.

[5] D. Holten, J. van Wijk. Force-directed edge bundling for graph visualization. Computer Graphics Forum, 28(3): 983–990, 2009.

[6] C. Dunne and B. Shneiderman. Motif simplification: improving network visualization readability with fan, connector, and clique glyphs. In CHI '13: Proc. SIGCHI Conference on Human Factors in Computing Systems: 3247-3256, 2013.

[7] L. Royer, et al. Unraveling protein networks with power graph analysis. PLoS Comput Biol 4(7): e1000108, 2008.

[8] P. Rosenstiehl, R.E. Tarjan. Rectilinear planar layouts and bipolar orientations of planar graphs. Discrete Comput. Geom. (1): 343-353, 1986.

[9] R. Tamassia, I.G. Tollis. A unified approach to visibility representations of planar graphs. Discrete Comput. Geom. (1): 321-341, 1986.

[10] B. Blakley. Reduction of Flow Diagrams to Unfolded Form Modulo Snarls, YLYK, Ltd. (1987) Final Report to AFOSR on Contract F49620-86-C-0103, Defense Technical Information Center, 1987.

[11] A.J. McAllister. A new heuristic algorithm for the linear arrangement problem. Technical Report 99_126a, Faculty of Computer Science, University of New Brunswick, 1999.

[12] J. Rumbaugh, I. Jacobson, G. Booch. *The unified modeling language reference manual*, *Volume 1*. Reading, MA: Addison-Wesley, 1999.

[13] A. Bezerianos et al. GeneaQuilts: A System for Exploring Large Genealogies, IEEE Transactions on Visualization and Computer Graphics 16(6):1073-1081, 2010.

[14] H. Salgado et al. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Research 41(D1): D203-D213, 2013.

[15] "Data Contest", *BioVis 2014*, http://www.biovis.net/year/2014/info/contest_data, 2014.

[16] S.M. Smith, et al. Resting-state fMRI in the Human Connectome Project. NeuroImage 80:144-168, 2013.

[17] J.R. Clough, et al. Transitive reduction of citation networks. Journal of Complex Networks 3(2):189-203, 2015.

[18] A. Kolliopoulos. The 1-Hyperbolic Projection for User Interfaces http://www.dgp.toronto.edu/~alexk/hyperproj.pdf, 2003.

[19] "HCP Citations", Human Connectome Project, http://www.humanconnectome.org/documentation/citations.html, 2014.